DIG64 Technical Whitepaper: IA-64 Server Network Communications Solutions

Intel Corporation
January 2000

Overview

This whitepaper discusses some of the networking and communications technologies available for high-end computing systems, and recommends which technologies should be incorporated into IA-64-based servers. This whitepaper is one of a series of technical whitepapers supporting the *Developer's Interface Guide for IA-64 Servers* (DIG64).

This document presents information developed by DIG64 subgroups and is offered to industry developers. However, this document does not specify DIG64 compliance requirements. See the *Developer's Interface Guide for IA-64 Servers* for the complete compliance requirements. This whitepaper assists in selecting and configuring network communications solutions for IA-64 servers to ensure usability and to provide reliable, fast network communications solutions for IA-64 server customers.

This whitepaper discusses issues for the following network communication devices:

- Network adapters
- T1 Interface
- ATM requirements

Network Adapter

A network adapter is a network interface card that connects to the server's I/O bus, providing a network interface for the server. This whitepaper does not prescribe the physical or data link layer of the network, and most of the network adapter features described are generic. Many requirements described here apply to other network communications devices such as ISDN, cable modem, and ADSL, which are described later.

With the ever increasing data throughput requirements for high-speed, IA-64-based servers, an ISA-based network adapter is inappropriate. Because IA-64 microprocessors are used in extremely high-speed servers, the network interface for these servers is a critical component of the platform. The network adapter should provide a high-speed physical network interface and should also off-load, to the fullest extent possible, network protocol tasks. This section describes features and capabilities that should be supported by the network adapter in IA-64 systems.

Because offloading provides better use of the primary processor resources, network adapters should support it. The following are examples of networking protocol tasks that can be off-loaded: computation of TCP/UDP/IP checksums, encryption of packets in compliance with standards such as IPSec, and message segmentation.

By default, the network adapter must negotiate full duplex operation when both the network adapter and switch port in a link-pair support full duplex, and the networking technology provides a standard way for each to detect and/or negotiate the duplex mode. Half duplex may be used if it is the only mode supported by one or both link partners, or if it can be manually configured when special conditions warrant it. The goal is to configure this setting automatically without end-user intervention and then to gain the best possible performance.

Where the network media allows it, the network adapter dynamically determines whether it is connected to a functional link partner such as a hub, switch, or router. The device should indicate the link state in the following cases:

- At boot time.
- After returning to the ACPI D0 power state.
- When the link state changes while in the D0 power state (no time limit is specified for the required detection or status indication).

If the adapter is on an expansion card and is not used as a boot device, then the device drivers can determine the presence of the functional link. If the device is not connected to a functional link partner, the driver must provide appropriate status indication.

As long as the particular network media employed allows it, network adapters that support multiple transceivers must be capable of automatically detecting which transceiver type is connected to the network. The network adapter must then automatically drive the correct connection. In all cases, the user must not be required to set jumpers, or to manually enter information informing the operating system of the transceiver type.

Buffer alignment refers to whether a buffer begins on an odd-byte, word, double word, or other boundary. Network adapters must be able to transmit packets where fragments are on an odd-byte boundary. For performance reasons, packets should be received into contiguous buffers on a double-word boundary. All communications must be free of errors across any bus bridge.

It is strongly recommended that server network adapters support the remote new system setup capabilities defined in *Pre-Boot Execution Environment Specification, version 2.1* or later. Also, to maintain server security, it must be possible to enable and disable the remote boot (remote new system setup) capabilities through administrative control.

Multiport network adapters should supply remote system setup capabilities on zero, any, or all ports.

To support efficient data transfer, network adapters interfaced to the server via the PCI Bus must support higher-level PCI commands. These are described in the *PCI Local Bus Specification, Revision 2*.

"Promiscuous mode" allows special applications/nodes to observe and analyze the network traffic for diagnostic and informational purposes, even if the traffic is not specifically intended for that application. This is a valuable traffic analysis tool. Network adapters must support operation in promiscuous mode. In this mode, the adapter does no address filtering. All network traffic on the network connection is reported to the software interface within the limits of the overall system, including software and hardware. This requirement applies only to LAN (nonswitched) media. By default, promiscuous mode is not enabled but may be enabled and disabled by software.

For IA-64 servers, the network adapter and driver need to support multicast promiscuous mode. By supporting this feature, the adapter and the driver enable performance improvements for special-purpose servers and applications, such as multicast routers. This requirement applies to those networking technologies that support multicast, such as Ethernet. It does not apply to technology, such as Token Ring, that does not support multicast. By default, multicast promiscuous mode is not turned on.

Quality of service (QoS) components provide link layer priority information to drivers on a per-packet basis. Priority values are derived by mapping IETF Integrated Services (intserv) service types to 802.1p priority values (referred to as the "user priority" object defined in *Integrated Service Mappings on IEEE 802 Networks*). The intserv service type used for the mapping is determined by QoS-aware applications and by QoS-aware operating system components.

To improve the system performance, PCI Bus-based network adapters must be capable of bus mastering. This allows the adapter to perform DMA operations, thereby decreasing the load on the system processor.

USB network communications device vendors should implement forthcoming networking extensions to the USB Class Definitions for Communications Devices. Vendors are also encouraged to participate in the definition and implementation of both USB and IEEE 1394 working-group efforts.

The system should awaken from a lower-power state based on network events specified by the local networking software. The result of a "wake-up event" is that any standard network access—such as connections to shared drives, and service and management applications—can awaken a system from lower-power states transparently.

The following network communications devices and their associated drivers should include support for wake-up events:

- Ethernet and Token Ring network adapters.
- Integrated DOCSIS cable modems.
- Other or future devices that transfer 802.3 Ethernet framed packets.

As defined in *Network Device Class Power Management Reference Specification Revision 1.0*, a network adapter and its driver should support wake-up on receipt of a network wake-up frame. Support for wake-up on detection of a change in the network link state or on receipt of a Magic Packet* event is optional.

To enable Wake on LAN* capability for basic networking scenarios, the network interface card should be capable of storing information describing a minimum of five wake-up packet patterns. The card also must be able to recognize wake-up packets based on pattern matches anywhere in the first 128 bytes of the packet. Better still, network adapters should be capable of storing information describing at least eight wake-up packet patterns to enable more advanced applications such as Wake on LAN capability on multiple home systems or on receipt of multicast packets, in addition to the above basic scenarios.

PCI-based network adapters should support the generation of a power management event (PME# assertion) from the D3 cold device state if the physical layer technology is capable of operating under the voltage and current constraints of the D3 cold device state. For example, 100baseTX adapters can meet this requirement based on the state of the art in mid-1998. 1000baseSX or 1000baseLX (gigabit Ethernet using optical fiber media) products cannot meet this requirement because of the power required to operate the optical physical layer.

For information regarding services and functionality discussed in this section, see the specification *Advanced Configuration and Power Interface (ACPI) Revision 2.0*.

T1 Interface

In this discussion of T1 implementation guidelines for IA-64 servers, "or" means that at least one of the options must be implemented, while "and" means that each of the options must be implemented. T1 interfaces are not required on IA-64 servers. If they are implemented, they need to meet the guidelines indicated in this section.

T1 interfaces are a good fit for server applications. They support full duplex user-data transmissions of up to 1.536-Mbps over a two-pair (4-wire), usually shielded copper wire running at a constant wire speed of 1.544-Mbps. The technology is well established and uniformly implemented across T1 service providers.

NOTE: T1 is a physical layer transport technology only. Effective use of the transport requires support of higher layer services.

For the integrated T1 interface, the network adapter requirements listed next must be met. For more information see "Network Adapter" section earlier in this whitepaper:

- Adapter automatically senses presence of functional network connection.
- Adapter can transmit packets from buffers aligned on any boundary.
- Adapter communicates with driver across any bridge.
- Device Bay network adapter meets requirements.
- PCI network adapters are bus masters.
- USB or IEEE 1394 network device complies with related device class specifications.

The eight-position modular jacks defined by USOC codes RJ48C and RJ48X are the only connectors acceptable for this platform. The RJ48C is recommended over the RJ48X. This guideline specifically excludes using a T1 device that presents only an unstructured interface at the connector.

Structured full or fractional T1 service must be provided (from 1 to 24 channels for user data).

Both Alternate Mark Inversion (AMI) and AMI with Bipolar Eight Zero Substitution (B8ZS) line coding formats must be provided as selectable configuration options. AMI and B8ZS are line coding formats used to eliminate a condition called "loss of data stream synchronization." Loss of synchronization occurs when the data stream consists of a series of consecutive ones and the receiver is confused as to which bit is which in the stream. B8ZS is preferred where available, but it may not be available in some locations or from some T1 service providers.

Timing modes must include network loop timing or internal timing. Both the loop timing and internal timing network modes must be selectable. A T1 receiver samples the line voltage to determine whether a bit is a one or a zero. If the clock pulse (the signal that tells a receiver to sample the line) drifts from the clock that was used to send the signal, data will be incorrectly interpreted at the receiver. For this reason, there is a preferred clock mode for T1 terminating equipment. The preferred clock mode synchronizes the clock of the terminating equipment with that of the sender. This is called slave, or *loop timing*. In this mode, the interface derives its clock from the received signal itself. In the United States, there is a clocking hierarchy in which the most accurate clock feeds down to all branches in the network. A device such as an IA-64 server will generally be a leaf node on this network.

Certain configurations exist where the IA-64 server drives a T1 network that is either permanently isolated from, or has temporarily lost connectivity with, the master clock hierarchy. In this case, the server's T1 interface and its connected equipment cannot be allowed to slave time to each other but must operate in internal timing mode where one of the endpoints assumes the role of clock master. An internally clocked IA-64 server provides the master clock for the T1 interface, and the remote device derives its clock from the master clock.

Where the T1 facility does not support B8ZS line coding, there is a restriction on the data patterns that can be sent to the line. The restriction is that the data stream must contain a minimum density of ones, so the T1 receiver can detect sufficient voltage transitions to synchronize the receiver. The interface is then running in restricted channel mode. However, constraining the data payload of an IA-64 server T1 interface client protocol is unacceptable.

This problem is resolved by setting a fixed bit in each DS0 channel to a logical one, satisfying the ones density requirement for T1. This means that only seven bits out of eight are available for user data, or 56-Kbps of each 64-Kbps DS0 channel. Configuring the DS0 data rate to 56K on T1 facilities that use AMI line coding ensures that all WAN data is transported cleanly on the T1. For this reason both 56-Kbps and 64-Kbps channel service must be selectable.

Another solution to the ones density problem is to invert the WAN data stream when HDLC-based services are supported. In this case, normal (nonaborting) data streams have at most six contiguous ones and will satisfy the ones density restriction when inverted. The abort condition may also be represented as exactly seven consecutive ones (e.g., 011111110). Accordingly, timeslot data can be inverted or normal. This selection is mandatory for all T1 interfaces that transport only HDLC-based data.

The T1 interface must be configurable to support any of the following Line Build Out selections.

Table 1

DSX-1 Line Build Out (short haul)	0 to 133 feet
	133 to 266 feet
	266 to 399 feet
	399 to 533 feet
	533 to 655 feet
DS-1 Line Build Out (long haul)	0 dB
	-7.5 dB
	-15 dB
	-22.5 dB

The IA-64 servers' T1 intrafStandard Framing (SF) and Extended Super Framing (ESF) patterns must be supported. To assist in synchronizing devices on a T1 line, frames containing the payload data of twenty-four 64-Kbps (DS0) streams are organized into a superframe. The addition of one bit per frame (for 192+1 = 193 bit/frame) allows a unique framing pattern to be inserted and detected; the pattern is read across multiple frames. The framing pattern thus occupies a bandwidth of 8-Kbps. SF formats have evolved, but SF now generally refers to a D4 format transmission, which is backward-compatible with the older D2 and D3 formats.

In the Extended Super Frame format, the 8-Kbps super frame bandwidth is partitioned into a 2-Kbps framing pattern (uniquely recognizable), a 4-Kbps data link, and a 2-Kbps cyclic redundancy check CRC. The CRC provides error-checking over the entire T1 data stream, and the data link allows for an embedded out-of-band control and reporting protocol. Loopbacks can be commanded, and remote statistics can be collected using the data link. The ESF format provides more functionality than the standard super frame format and is preferred for IA-64 servers' T1 interfaces.

Maintenance Signals for SF & ESF formats

- ANSI T1.403 PRM Generation and Reception
- ANSI T1.403 Payload and Line Loopback Activation and Release
- AT&T TR54016 On-demand PRM Response Generation
- AT&T TR54016 Loopback Activation and Release
- AT&T TR54016 Alarm Handling

Asynchronous Transfer Mode (ATM)

Asychronous Transfer Mode (ATM) is not required for IA-64 servers, however, if it is implemented, the ATM adapter must meet the guidelines defined in this section. For more information, see *ATM User-Network Interface Specification, Version 3.1*.

Asynchronous Transfer Mode (ATM) has been used in the industry since 1990. It allows users to seamlessly generate a flow of data between local area networks (LAN) and wide area networks (WAN). Data generated by a PC flows through ATM LAN technology to an ATM switch or router point on the same LAN. From this ATM switch or router, data is sent through either a Permanent Virtual Circuit (PVC) or Switched Virtual Circuit (SVC) and arrives at a distant location on the ATM backbone. In turn, ATM integrates LAN and WAN resources.

ATM uses fixed size packets, called cells, to transfer information. Each cell is exactly 53 bytes long. Of those, 5 bytes contain header, address, and descriptor information. The other 48 bytes contain the information being transferred across the ATM network. ATM determines the destination of data by using the User-to-Network Interface (UNI) and the Network-to-Network Interface (NNI) information in the ATM cell header. At each network node, buffers are assigned at both the transmit and receive locations of the virtual connections.

ATM sets up the actual circuits ahead of time by using a predetermined mapped network to work with. The following three main characteristics make each virtual connection unique in an ATM system:

1. Either a PVC or SVC comprises the trunk that the virtual connection resides on.
2. A Virtual Path Identifier (VPI) and a Virtual Channel Identifier (VCI) are assigned at various nodes throughout the network and to the VPI and VCI assigned at the User-to-Network Interface (UNI) points.

3. The cell speed or rate accepted over the VC in question.

ATM adapters must meet the network adapter requirements listed in the DIG64 specification. The following network device requirements also must be met:

- Adapter automatically senses the presence of a functional network connection.
- Adapter can transmit packets from buffers aligned on any boundary.
- Adapter communicates with a driver across any bridge.
- Device Bay network adapter meets requirements.
- PCI network adapters are bus masters.
- USB or IEEE 1394 network devices comply with related device class specifications.

ATM adapters also must support a minimum number of simultaneous connections. The VPI and VCI ranges supported by the adapter affect the maximum number of simultaneous connections supported on a system.

This requirement affects the applicability of the adapter to ATM applications such as LAN Emulation, where at least one dedicated virtual channel is created between each pair of communicating ATM hosts.

Table 2

System Type	Simultaneous Connections
Client (ATM adapter)	64 or more
Client (Integrated ATM/ADSL-adapter)	16 or more
Server	2048 or more

The ATM adapter supports all service types defined by the ATM Forum. The ATM adapter should support the constant bit rate (CBR), variable bit rate (VBR), available bit rate (ABR), and unspecified bit rate (UBR) service types as defined by the ATM Forum.

CBR is a reserved bandwidth service. A contract is established between the network and the end station. The end station provides the network with parameters describing the traffic for that specific connection at call setup time. The network uses call admission control, which allocates resources that match the parameters, or if the resources are not available, rejects the call. Once the call is accepted, it is the end station's responsibility to send only traffic that is compliant with the contract. The network checks the traffic against the contract, and noncompliant cells are discarded. Support for at least two simultaneously active VBR or CBR connections is required for basic ATM signaling and management. This service is used for emulating circuit switching. The cell rate is constant with time. CBR applications are sensitive to cell-delay variation. Examples of applications that can use CBR are telephone traffic (i.e., nx64-Kbps), videoconferencing, and television.

Like CBR, VBR is a reserved bandwidth service. The network allocates resources at call setup in response to the traffic parameters requested by the end station. However, in the case of VBR, in addition to a peak rate, a sustainable rate and a maximum burst size are established. The sustainable rate is the upper limit of the average rate, and the maximum burst rate limits the duration of cell transmission at peak rate. These additional parameters allow the network to achieve statistical multiplexing by allocating fewer resources for the connection than would be required by the peak cell rate. This service is designed for applications that are sensitive to cell-delay variation. Examples of real-time VBR are voice with speech activity detection (SAD) and interactive compressed video.

An ATM adapter must support a minimum number of simultaneously active VBR or CBR connections. Both connections are well suited for supporting applications with stringent requirements for quality of service (QoS), such as multimedia transmission or high-quality videoconferencing.

Support for more VBR/CBR connections is needed for ATM adapters that support multimedia or other traffic that demands QoS. Such system types are shown in the following table:

Table 3

System Type	Simultaneous active VBR/CBR connections
Client	6
Server	500

An available bit rate service (ABR) provides rate-based flow control and is aimed at data traffic such as file transfer and e-mail. Although cell transfer delay and cell-loss ratio don't have to be guaranteed or minimized, it is desirable for switches to minimize delay and loss as much as possible. Depending on the state of congestion in the network, the source is required to control its rate. Users are allowed to declare a minimum cell rate, which is guaranteed to the connection by the network.

An Unspecified Bit Rate (UBR) is a nonreserved bandwidth service. The service responds to congestion by dropping cells when their buffers become full. The cell loss ratio is unspecified, which means that the network is not required to provide resources for a proposed UBR connection. No flow control parameters are specified in the ATM Forum for UBR service. Applications send data across the network with no guarantee when or if that data will arrive at its destination. However, a USB service is useful for all applications without real-time characteristics. UBR is used by default for standard ATM services such as LAN Emulation and IP over ATM. In addition, Point-to-Point Protocol (PPP) is widely used for residential network access, and UBR is used by default for PPP over ATM virtual circuits. Therefore, ATM adapters must support the UBR service type.

Traffic shaping is a mechanism for using and enforcing the quality of service (QoS) over an ATM network, and is based in part on the Generic Cell Rate Algorithm (CGRA). The CGRA Algorithm is called a virtual scheduling algorithm. The algorithm works by checking every cell to see if it conforms to the parameters for its virtual circuit. The CGRA has two parameters: the maximum allowed arrival rate (T) and the amount of variation that is tolerable (L). T is the reciprocal of the peak cell rate (PCR). L is the cell delay variation tolerance. T-L is the minimum boundary for two cells to be received. The ATM adapter must support and enforce all the traffic-shaping rules specified for each service type it supports, including CBR, VBR, ABR, and UBR. This requirement includes enforcement of peak-cell rate on UBR virtual circuits.

An ATM adapter must enforce a PCR on UBR virtual circuits. ATM adapters are used to connect router, remote access, and content servers to the public ATM network. High-speed residential broadband access networks such as Asymmetric Digital Subscriber Line (ADSL) and cable modem use an ATM virtual circuit from home or small office computers to connect directly to these servers.

To avoid packet loss and ensure efficient network use, it is critical that all ATM, integrated ATM/ADSL adapters, or ATM/cable modem adapters enforce requested PCR on UBR virtual circuits. Because any ATM adapter might be installed in a server to which clients connect through the public network, this guideline applies to all ATM adapters.

An ATM adapter must support dynamic link speed configuration. To support this feature, the PVC or SVC involved in the virtual connection is dynamically created. When connected to a residential broadband network, ATM adapters must restrict the aggregate transmission rate across all active virtual circuits so that the rate does not exceed the provisioned upstream bandwidth of the residential broadband network.

All integrated ATM/ADSL and ATM/cable modem adapters must support aggregate shaping of upstream bandwidth according to the provisioned or trained bandwidth, whichever is lower. Some implementations can support rate adaptation, and so lower-than-provisioned rates may be negotiated because of poor line conditions.

All 25-Mbps ATM adapters must support this as well, because any 25-Mbps ATM adapter could be used to connect through an external ADSL modem to an ADSL network. This support is optional for ATM adapters with line rates higher than 25-Mbps.

An ATM adapter must support Operations, Administration, and Maintenance (OAM). OAM is important for optimizing the functionality of the ATM adapter. Special cells provide information relating to cell flow. OAM cells can be sent on the network using a unique header that identifies the cell as an OAM cell. These cells have a limited frequency. OAM is required to perform:

1. Fault Detection.
2. Fault Localization.
3. Performance Monitoring.

At minimum, the ATM adapter must respond to received F4 and F5 loopback OAM cells. Support for other layers, F1–F3, is optional.

An ATM adapter must support buffer chaining (Tx + Rx). This feature is needed for large packets. This capability is required for server systems but is recommended for client systems.

Summary

This whitepaper is concerned with the selection and configuration of network communication solutions, for which interface specifications are published and widely available. The goal of this series of documents is to provide sufficient technical information to designers and integrators of subsystem components for use in IA-64 server systems so that the resulting systems are highly reliable, available, serviceable, and fully performant. It is not the goal of these documents to specify or define operating system policies, application usage characteristics, or the philosophy of the design of IA-64 server subsystem components or devices.

For a more in-depth discussion of each of these technologies, refer to the list of documents in the References section below. In this paper, no consideration has been given to incomplete or in-progress technologies or interface specifications. Future technologies will be addressed when those interface specifications are published and made available.

References

Advanced Configuration and Power Interface (ACPI), Revision 2.0
http://www.teleport.com/~acpi/

ATM User-Network Interface (UNI) Specification, Version 3.1
http://www.atm.org

Developer's Interface Guide for IA-64 Servers
http://www.dig64.org

IEEE 802.3 Ethernet Specification
http://www.ieee.org

IEEE 802.14 Media Access Control Protocol
http://www.ieee.org

IEEE 802.1 Standard for Local and Metropolitan Area Networks: Virtual Bridge Local Area Networks
http://www.ieee.org

Network Device Class Power Management Reference Specification, Revision 1.0
http://www.pcisig.com/

Pre-Boot Execution Environment Specification, Version 2.1
http://developer.intel.com/ial/wfm/wfmspecs.htm

RFC 1717 The PPP Multilink Protocol
http://www.ietf.org